Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 7382569, 12 pages
https://doi.org/10.1155/2020/7382569

Research Article

Using Clustering Methods in Multinomial Logit Model for Departure Time Choice

Shahriar Afandizadeh Zargari and Farshid Safari

Department of Transportation Engineering and Planning, School of Civil Engineering, Iran University of Science and Technology, Tehran 16846-13114, Iran

Correspondence should be addressed to Shahriar Afandizadeh Zargari; zargari@iust.ac.ir

Received 8 August 2018; Revised 3 January 2019; Accepted 18 August 2019; Published 15 January 2020

Academic Editor: Francesco Viti

Travellers have to make some decisions for each trip, and one of them is the choice of departure time. Discrete choice models have been employed as an approach to departure time modelling by many researchers. In this method, preparing choice set is a primary challenge which involves the definition of some departure periods to be selected by the traveller. In this research, choice sets were formed by applying the clustering methods on departure times. Afterwards, we developed Multinomial Logit (MNL) models on different choice sets and compared the models. The data used throughout this research belonged to Mashhad City. Research results indicated that Ward's hierarchical clustering method is improper for time discretization; furthermore, the K-means clustering method is more efficient than the expectation maximization and K-medoids methods in the time discretization for MNL modelling. The developed model (based on K-means clustering method) accurately predicts departure time for 58% of persons within the test group, which reflects the effectiveness of the resulting model compared to the 36% which is obtained without the model.

1. Introduction

Travellers' choice plays a substantial role in the performance of the transportation system; thus, models which will predict traveller behaviour and their choices are highly valuable. In other words, travel demand is the result of various decisions which are made by travellers and modelling of these choices is profoundly influential. One of the choices available in the travelling process is the departure time choice (DTC). It is essential to develop an appropriate DTC model due to the nature of departure time, its application, and complexity of time representation; also, developing an accurate model that is closely related to precise generation of choice set.

Firstly, based on an investigation on nature of traveller's decisions [1], the choice behaviour can be classified into the following three levels: strategic, tactical, and daily level. The DTC can be allocated to the tactical and daily levels; therefore, it is not a long-term choice and may be changed based on travellers' daily condition. Furthermore, the nature of DTC has unique characteristics due to travel time uncertainty and departure time can be affected by demand management policies.

Besides, the wide range of applicability of DTC is another issue that emphasizes the importance of modelling DTC. For instance, its application in peak hour estimation, evaluation of travel demand management policies, and dynamic traffic assignment is vital. The peak hour coefficient is usually used in four-step models to convert daily demand to peak hours [2]. In other words, time of peak hour demand is calculated with a constant coefficient (resulted from surveys), and the performance of transportation system is assessed for peak hours. But departure time models can help to reach better estimation from the share of demand in various time periods. Nonetheless, in assessing the effect of different management policies (including demand management policies), it is necessary to examine how travel demand is distributed in a day. The departure time influences the traffic peak hour and its pattern; therefore, developing a DTC model which is applicable to demand management policies is required [3–5]. Another primary application of these models is in dynamic traffic assignment studies because development of accurate departure time models improves the performance of these models [6]. A further application of the DTC is in activity-based models [7, 8].

The definition of choice set for DTC models is a complicated problem due to the following considerations which should be taken into account. First, the conversion of continuous time to discrete time and establishment of a rational relation between temporal resolution and model complexity must be considered. Second, the relationship between choices should be taken into account, especially in short periods. Third, the person's perception of choices depending on travel time must be considered. Many people round off the time, which depends on travel time and its variability. For instance, in short trips, 9 : 48 may be rounded to 9 : 50, while it may be rounded to 10 : 00 for long trips.

Moreover, the accuracy of DTC modelling is mainly influenced by choice set generation. The choice set generation refers to the definition of departure periods which are considered and selected by the person. Thus, to achieve an accurate DTC model using the appropriate method for choice set generation is inevitable.

After we generally stated the importance of DTC models, research motivation will be discussed in this section. There were two main motivations for this research which are firstly, to emerge a method for estimating the peak hour pattern based on estimated socioeconomic characteristics of travellers at planning horizon and secondly, developing an approach to evaluate the effect of travel demand management in four-step models.

In the first place, because of changeable socioeconomic attributes of travellers during years of planning period, the choice of departure time will be changed and the pattern of peak hours will be varied consequently. Therefore, using the observed pattern of peak hours in the base year (based on survey) for planning horizon would lead to inaccurate estimation of traffic volumes. Accordingly, we tried to make a DTC model based on observed behaviours so the peak hours can be predicted accurately using the DTC model and estimated socioeconomic variables of planning horizon.

In the second place, the analysis and evaluation of demand management policies (for example network pricing in a specific duration of time such as peak hours or making changes in starting and ending time of activities in administrative offices and educational centres) and their corresponding impact on DTC in the four-step models cannot be performed properly. Therefore, authors made an effort to make a DTC model and incorporated sociodemographic variables as well as trip attributes within the model. This model can be a tool to assess demand management implications on the departure time choice and consequently on the pattern of peak hours.

In this paper, we review modelling departure time and choice set generation in literature and examine a different method of clustering as a tool for choice set generation and compare the result of departure time modelling based on these various techniques. The DTC and choice set generation (as a component of DTC model) is performed in the city of Mashhad with a population of about 2.7 million people as a case study. This paper applied the multinomial logit (MNL) model to estimate the departure time choice for home-based work trips, which constitute a significant proportion of urban trips.

The remainder of this paper is organized as follows. First, we review the literature on DTC; then we provide an overview of the problem definition. In the next section we present methods. Finally, we emphasize the results and findings from this study respectively, and last section presents the conclusion.

2. Literature Review

Departure time models reflect the selection of a point or period at which the person begins the trip. The DTC models have been developed with various methods based on the framework of application. The main applications of this models are using DTC as a subcomponent of dynamic assignment or an element of activity scheduling modelling, using DTC as a tool for predicting peak hours pattern in long-term planning (as a time of day model) and using DTC as a tool for analysis and evaluation of demand management policies.

2.1. The DTC Model as a Subcomponent of Dynamic Traffic Assignment or an Element of Activity Scheduling Modelling. Lim et al. [6] investigated the logit-based combined departure time and dynamic stochastic user equilibrium assignment problem. In this study, for each OD pair whenever road users choose their departure times and route according to the logit choice model, considering time dependent travel cost and schedule delay.

Feil et al. [8] acknowledged that scheduling in activity-based travel demand modelling follows three major lines of research which are econometric models, utility-based microsimulations and computational process models (CPMs). Econometric models have some advantages and they are based upon a well-established statistical methodology. Utility-based microsimulations apply a sequential decision making process. In this approach, rather than a probability distribution, the result is always a precise solution alternative.

Adnan [9] developed an integrated model for scheduling the activities an individual is supposed to do in a given day with the representation of road network congestion effects. In recent works of activity scheduling modelling for a daily activity travel pattern, the utility specification of their model includes a component that measures the utility of activity engagement. This has been calculated through predetermined time-of-day-dependent marginal utility profiles for a particular activity.

Cantelmo et al. [10] considered the problem of jointly modelling activity scheduling and duration within a Dynamic Traffic Assignment (DTA) problem framework. These researchers introduced the travel choice model as a subcomponent of DTA which consists route, departure time, and mode choice. This study considered the final daily activity pattern is a function of travel time, activity duration and the preferred arrival time at the destination. The utility function that is used in activity scheduling modelling at this study contained the clock-based and duration-based utility.

2.2. The DTC Model as a Tool for Predicting Peak Hours Pattern in Long-Term Planning (as a Time of Day Model). Small and Biggiero [11, 12] employed MNL models to select a departure

time. In 1987, Small [13] proposed an ordered generalized extreme value model, in which m departure time periods were considered in a nested design. Hendrickson et al. [14] used the data collected from Pittsburgh (Pennsylvania) to examine flexibility of departure time choices for work trips. These researchers estimated the logit model for concurrent selection of transportation mode and departure time.

Chin believes that although many studies have been conducted on transportation modes, few studies have discussed DTC and thus the importance of this problem is not understood. This study was an attempt to model DTC using the MNL model and analyse the nested logit model. Research results [15] revealed that delay, travel cost, and travel time are among the significant factors influencing the DTC for commuters. Sumi et al. [16] carried out a study to predict the response of commuters to operational features of the public transportation system for departure time and route choice.

In these applications of the MNL model, a day is divided into some periods, and the MNL model is used to select a choice from the set of alternatives (periods). In the MNL model, systematic utility functions are defined as a function of socioeconomic attributes and variables related to trip purpose. Although the MNL model overlooks the similarities or correlations among adjacent periods and violates the IID assumption, it is still popular due to its closed form and extensive use [17].

In many urban areas, the four-step process will continue to be used for macroscopic modelling of large scale road network. It is typical for models to start by estimating daily travel in the trip generation step, but traffic assignment is performed separately for different time periods. Typically, a four-step model uses three to five periods (for example, morning peak, mid-day, afternoon peak, night). The most common method for conversion of daily demand to some periods in four-step models is simple factoring. These factors typically are developed from the temporal patterns of trips which attained from OD surveys. While this method is relatively easy to implement and to apply, it is not sensitive to varying transportation conditions, limiting its usefulness in analysing policy changes or congestion management activities, and these time coefficients aggregated from OD surveys are somewhat biased, thereby reducing accuracy [3].

Fujita et al. have proposed a model for estimation of these time coefficients using links volume counting and Semi-dynamic Traffic Assignment [2]. In that study, a model has been developed that justifies the 24-hour time coefficients under a given day-long OD demand by minimizing the least square error between hourly observed link flow and estimated link flow But this research is useful for operational analysis, not for strategic planning because the analyst does not have access to link volumes in horizon year.

2.3. The DTC Model as a Tool for Analysis and Evaluation of Demand Management Policies. Mannering et al. [18] examined the effects of provision of traffic information on commuters' behaviour in Seattle. Ettema et al. [19] developed a behavioural DTC model based on travelling activities plans. They used the utility of participation in activities to model the DTC. Ettema et al. [20] also considered a variation of departure time as a possible response to density in microscopic traffic models. He [5] studied the effect of access to a flexible work plan on commuters' choice of departure time. Results of their investigations revealed that people with flexible work plans leave later.

2.4. Various Econometrics DTC Models. Different econometrics models used for selecting the departure time include MNL model, nested logit model, cross-nested logit model, mixed logit model, continuous time mode, and ordered generalized extreme value model (OGEV).

Bhat [21] referred to the importance of the travel mode and departure time choices and stated that although the priority of researchers modelling travel demand was travel mode choice, little attention has been paid to the DTC. He introduced a nested structure by considering the travel mode choice at the higher level and the DTC in the lower level. The MNL model was used for travel mode choice, while the ordered generalized extreme value (OGEV) model was used to consider ranks of choices and selection of departure time. In another article, Bhat [22] introduced the mixed logit model for modelling the travel mode choice and departure time in home-based social-recreational trips. The relationships between alternatives can be considered in the cross-nested logit model by placing choices in several nests and omitting independence of subsets [23]. A structure that depends on the cross-nested logit model is the continuous cross-nested logit model, which is employed when the day can be divided into many short periods and all of the possible correlations can be considered in this modelling structure. Lemp et al. [24] used the same approach, which is highly reliable but does not allow the probability function to maintain its closed form. These researchers used the Bayesian estimation technique and reported that the model performs similarly to the continuous logit model. Their explanation of the continuous logit model is based on an application of the MNL method for a large number of discrete time choices.

Steed and Bhat [25] examined DTC for nonwork trips. The discrete choice model was estimated for shopping home-based trips and social/recreational home-based trips. Results indicated that the departure time of social/recreational trips and shopping trips is not highly flexible and is limited to certain times due to the person's or household's time limitations. In another study, Bhat and Steed [4] used the "hazard-based duration model" to develop a continuous model of DTC for urban shopping trips. Application of the continuous time model in predicting time variations in shopping trips is shown concerning changes in sociodemographic properties as well as the trip chain behaviour. Habib et al. [7] used the "hazard-based duration model" to select the departure time. Application of this model for DTC does not necessarily reflect behaviours. In such cases, the departure time is only modelled as a nonlinear regression model, and the behavioural trade-off is not reflected in the selection of departure time.

Jou et al. [26] used the reference point hypothesis of the prospect theory in selecting the departure time. Results indicated that according to the prospect theory, travellers display asymmetrical responses to loss and gain.

Ben-Elia et al. [27] proposed a DTC model based on the latent preferential entrance time notion. Using this model, they developed a DTC modelling framework by assuming a latent class of the preferred entrance time.

Sasic et al. [28] proposed a discrete choice model with a latent choice set. This model is ranked with generalized extreme value (GEV) models. The DTC model was developed for home-based commutes, and separate models have been developed for the departure time of users of private cars and public transportation.

2.5. Remarks and Conclusion. We did not find appropriate DTC model that is useful on four-step strategic planning. The proposed model in the study of Fujita et al. [2] is useful for operational analysis, not for strategic planning because the analyst does not have access to link volumes in horizon year. Ettema [29] suggested that sociodemographic variables be incorporated within the DTC model. We developed a DTC model and incorporated sociodemographic variables as well as trip attributes within the model.

The departure time choice and activity scheduling modelling are inextricably bound up, but it should be noted that the utility function in activity scheduling is derived from activity engagement (from clock-based MU and duration-based utility). In these models alternatives are activities. On the other hand, in many researches with econometric models the alternatives are time intervals.

The proposed DTC model in our paper can be used for determining the peak hour pattern at the horizon of long-term planning. However, it is essential to predict socioeconomic variables that incorporated in the model for planning horizon. Furthermore, this model can be a tool for assessment of demand management implications on the departure time choice and consequently on the pattern of peak hours.

3. Materials and Methods

3.1. Problem Definition. Several approaches to the DTC modelling are available in the research literature. This study investigates this problem using discrete choice models. Time discretization (formation of choice set from which the person picks an alternative) is particularly important in this approach. In this research, discretization was carried out using different clustering methods, and the MNL models were built based on a choice set including these clusters. In other words, the MNL modelling procedure is repeated for each clustering type, and the resulting models are compared. Although clustering and data mining have been employed in traffic engineering and transportation planning [30–32], no study has investigated the application of clustering in discrete choice. In the following, a brief description of MNL model is presented followed by general notions of clustering.

3.2. Multinomial Logit Modelling. In the discrete choice models, each choice is favourable to a certain degree for a person, and the alternative with the highest level of utility will be selected. Equation (1) determines the utility function of the person (n) that corresponds to alternative i in the choice set C_n.

$$U_{in} = V_{in} + \varepsilon_{in}, \tag{1}$$

where, V_{in} is the deterministic term for the utility and ε_{in} is the random term of the utility, which denotes the uncertainty caused by the limited power of the analyst. In the logit model, it is assumed that the random term has a Gumbel distribution [17]. Since the alternative with higher utility will have a higher chance of being selected, the probability of selecting choice i by the person (n) from the choice set C_n equals:

$$P(i|C_n) = P\left[U_{in} \geq U_{jn} \forall j \in C_n\right] = P\left[U_{in} = \max U_{jn}\right], \tag{2}$$

$$P(i|C_n) = P\left[\varepsilon_{jn} \leq \varepsilon_{in} + V_{in} - V_{jn} \forall j \in C_n\right]. \tag{3}$$

If ε_{in} is known, $P\left[\varepsilon_{jn} \leq \varepsilon_{in} + V_{in} - V_{jn}\right]$ is the cumulative distribution function of ε_{jn}, and considering the independence of ε values, this probability equals the multiplication of the cumulative distribution by all choices $j \neq i$.

$$P(i|C_n)|\varepsilon_{in} = \prod_{j \neq i} e^{-e^{-\left(\varepsilon_{in} + V_{in} - V_{jn}\right)}}. \tag{4}$$

Since ε_{in} is unknown, the probability of selecting choice i will equal the integral of $P_{in}|\varepsilon_{in}$ using all values of ε_{in}.

$$P(i|C_n) = \int \left(\prod_{j \neq i} e^{-e^{-\left(\varepsilon_{in} + V_{in} - V_{jn}\right)}} \right) e^{-\varepsilon_{in}} e^{-e^{-\varepsilon_{in}}} d\varepsilon_{in}. \tag{5}$$

This integral can be written in the following closed form [33]:

$$P(i|C_n) = \frac{e^{V_{in}}}{\sum_j e^{V_{jn}}}. \tag{6}$$

The deterministic term of the utility, V_{in}, for each choice is a function of the features of that choice and characteristics of the person (n).

$$V_{in} = V(z_{in}, S_n, \beta), \tag{7}$$

where z_{in} is the vector of qualities of choice i perceived by the person (n) and S_n is the vector of the person's characteristics, and β is the vector of utility function's parameters. The deterministic term of the utility is determined in calibration.

For the estimation of the utility function of the logit model, the probability of selecting a choice by the person (n) is expressed by $\prod_i (P_{ni})^{y_{ni}}$. If the person ($n$) selects choice i, then $y_{ni} = 1$; otherwise, $y_{ni} = 0$. Assuming that the choice of each person (n) is independent of the others, the probability of selecting an observed choice by each person is equal to:

$$L(\beta) = \prod_{n=1}^{N} \prod_i (P_{ni})^{y_{ni}}, \tag{8}$$

where β is a vector including the model parameters, and Equation (8) in logarithm form is as follows:

$$LL(\beta) = \sum_{n=1}^{N} \sum_i y_{ni} \ln P_{ni}. \tag{9}$$

β is estimated by maximizing this function. In 1974, McFadden indicated that $LL(\beta)$ is a concave function for a utility function with linear parameters, and numerous statistical software products are available for estimating these models [33]. In this research, the MNL models were estimated in Nlogit 5.0 [34].

3.3. Clustering. Clustering refers to the process of grouping data into clusters, as a result the similarity between the data inside each cluster is maximized, and the similarity between different clusters is minimized. Different clustering methods include the following: partitioning, hierarchical, density-based, grid-based, model-based, and constraint-based methods and large data clustering methods [35].

We will briefly describe the four clustering methods employed in this research. Firstly, k members are randomly selected which are assumed as the clusters' mean in the K-means method. Then each object is assigned to a cluster which is the most similar, based on the distance between object and cluster's mean. Afterwards, a new mean value is calculated for each cluster. Frequently, the SSE measure, which is defined as follows, is used to assess similarity [35].

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2. \tag{10}$$

In Equation (10), E is the sum of squared errors of all the dataset members, p is a point in the space and denotes a member, and m_i stands for the mean of cluster C_i. The K-means clustering method is sensitive to outliers because these points can deform the distribution form. For solving this problem, the k-medoids method is employed, and instead of considering the mean values of members of a cluster to be the reference point, one member is assumed to be the cluster's representative. Each member is assigned to a cluster with the most similarity to the cluster's representative. Members are partitioned based on minimization of the sum of dissimilarities between them and their cluster reference point. Therefore, the absolute error value is defined as follows:

$$E = \sum_{j=1}^{k} \sum_{p \in C_j} |p - o_j|. \tag{11}$$

In Equation (11), E denotes the sum of absolute error values for all dataset members, p is a point in space and indicates a member of cluster C_j, and finally o_j represents cluster C_j.

The hierarchical clustering methods group data into a tree of clusters and these methods can be classified into two groups: agglomerative and divisive. The weakness of hierarchical clustering methods is that they are unable to apply modifications after splitting, merging, or forming a new cluster.

Model-based methods maximize the fit of the existing data to some mathematical models. These methods are mainly based on the assumptions that the data is obtained by combining several probability distributions, a parametric probability distribution can display each cluster, and the data are a combination of these distributions. According to these assumptions, the goal is to estimate distribution parameters to allow for the optimum fit of parameters with the data. The EM method is an algorithm that searches for estimates of probability distribution parameters during iterations. The algorithm improves the parameters (clusters) by going through the following steps iteratively:

3.4. Step 1 (Expectation). Assign each member, x_i, to cluster C_k with the following probability.

$$P(x_i \in C_k) = p(C_k|x_i) = \frac{p(C_k)p(x_i|C_k)}{p(x_i)}, \tag{12}$$

where, $p(x_i|C_k) = N(m_k, E_k(x_i))$ has a normal distribution with a mean of m_k. In other words, this step calculates the probability of membership of x_i in each cluster.

3.5. Step 2 (Maximization). Estimates of model parameters are improved using the estimates of probabilities resulting from the previous step. For instance,

$$m_k = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i P(x_i \in C_k)}{\sum_j P(x_i \in C_j)}. \tag{13}$$

Since the present research has focused on the DTC modelling, clustering was only carried out on departure time, and from the methods mentioned above, two partitioning methods (K-medoids, K-means), one hierarchical approach, and one model-based method have been used.

4. Case Study

The data used for this research were obtained from Mashhad city which is the second largest city of Iran. This data included travel information and socioeconomic characteristics of travellers and were collected by interviewing 1.56% of the population. Its first O/D survey was conducted in 1994 and in 2008 was updated with another O/D survey, when it had a population of over 2.7 million. Figures 1 and 2 illustrate some descriptive characteristics of sample and Figure 3 shows the temporal distribution of trips in a day which aggregated by gender.

The trips under study included commutes aimed at work or education between 2 A.M. and 13 P.M. It was also assumed that travellers below the age of 10 obey their parents' decisions, and thus only the information of persons older than ten was used in this research. While the total number of trips was recorded to obtain more than 12000 with details, we carried out modelling with 70% of the existing data, and the remaining 30% was used to validate the model.

5. Results

In clustering, it is essential to determine the number of clusters, because it influences the results, so we considered the same number of clusters in all clustering methods. The EM (expectation-maximization) method resulted in six clusters, and the scree plot (which is mostly used with the K-means method) also confirmed six clusters. Figure 4 shows the sum of squared errors (SSE) inside clusters versus the number of clusters. Hence, the number of clusters was assumed to be six. We used the R software [36] in this research to cluster the data.

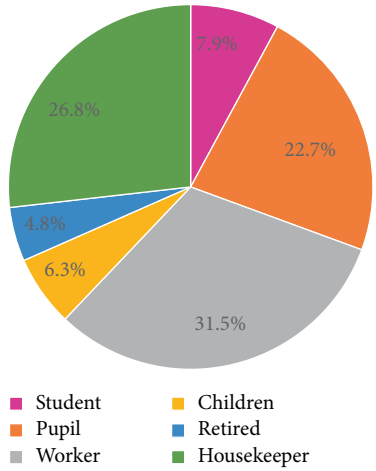

FIGURE 1: Share of employed and unemployed population of Mashhad city.

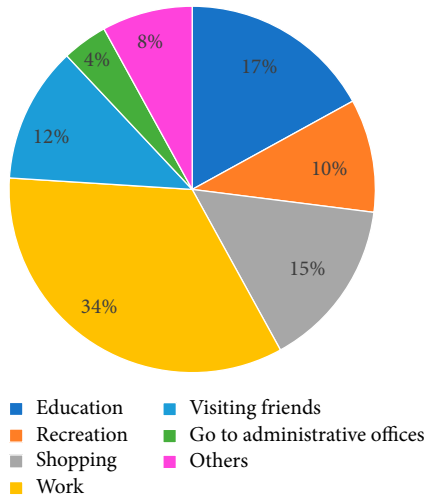

FIGURE 2: Share of different trip purposes from the total trips in 24 hours.

Clustering of the data with the four methods led to the formation of different clusters. The start and end time of each cluster in different clustering methods are presented in Table 1. Figure 5 illustrates data scattering from 2 A.M. to 13 P.M. Clusters are also shown by different colours and symbols in this chart. As it can be seen, the K-medoids method divided the peak period into more clusters, and clusters with shorter length formed in the peak hours. Therefore, the variability of the number of each cluster member is smaller than other methods (see Figure 6).

Besides, the first cluster in the EM method is smaller than the first cluster of others, which implies other characteristics of pre-peak-hour trips. In the K-means method, the duration of different clusters (except for the first cluster) is close to each other, but this method displays the highest standard deviation and range of the number of members.

Figure 5 depicts the overlap of clusters 2 and 1 (the red triangle with a black circle), clusters 3 and 2 (green plus and red triangle), and clusters 6 and 5 (pink triangle and turquoise rhombus) in the clustering carried out by the hierarchical method. The independence of choices and no overlap are requirements for MNL models which are not met by the hierarchical clustering method. Hence, we did not use the hierarchical clustering method in MNL modelling.

After discretizing time with the clustering as mentioned earlier, the discrete choice modelling is carried out. First, a choice set should be defined, and in this research, the rivals of each choice are its adjacent alternatives. In other words, the decision-maker starts the trip earlier or later than one or two periods, and this is because of the sequential nature of choices. We considered the following two states, and a schematic view of these states is depicted in Figure 7:

(1) Defining the choice set as a 6-membered set (see Figure 7(a)). It means all periods are feasible for travellers to start the trip.

(2) Defining the choice set as a 2-membered set (see Figure 7(b)) or 3-membered set (see Figures 7(c)–7(e)). A choice set with two members is for beginning and ending alternatives which means a traveller, who chooses the beginning or ending period, compares it with the first next or prior period. A choice set with three members is for the middle alternatives, which conveys a traveller, who chooses the middle period, compares the middle period with one previous and one next period.

The MNL modelling was carried out for two states (the 6-membered and 2-/3-membered sets). We examined different combination of independent variables to form the utility functions based on statistical indexes and our intuition. But we presented only the best developed model in Table 2, S1 and S2.

Results of model validation revealed that use of the 2-/3-membered sets in the DTC modelling yielded better than 6-membered sets. Therefore, we estimate the utility functions' coefficients of the MNL models for the 2-/3-membered sets of different clustering methods. Table 3 presents some criteria of MNL modelling based on various clustering methods, and Table 2 shows the utility functions' coefficients for modelling on choice set resulted from K-means method and their corresponding values of the *t*-test (See Tables S1 and S2 in the Supplementary Material for the modelling on the choice sets resulted from K-medoids and E.M. methods). In MNL modelling, variables namely gender, age, profession, vehicle type, vehicle ownership, education level, and travel time are significant (See Table S3 in the Supplementary Material for the description of variables that used in modelling).

As we expected, results of coefficient (parameter) estimation which is illustrated in Table 3 indicate traveller's job has a significant effect on departure time choice. This is predictable because of different start time and activity duration for various jobs. Moreover, the effect of selected mode on departure time choice is unsurprising. The public transportation users are more likely to depart in a period between 7.28 and 8.30. The pedestrians and bicycle riders are more reluctant to depart before 6.17 than the period between 6.17 and 10.75, and they are more prone to leave their homes between 10.75 and 13.00.

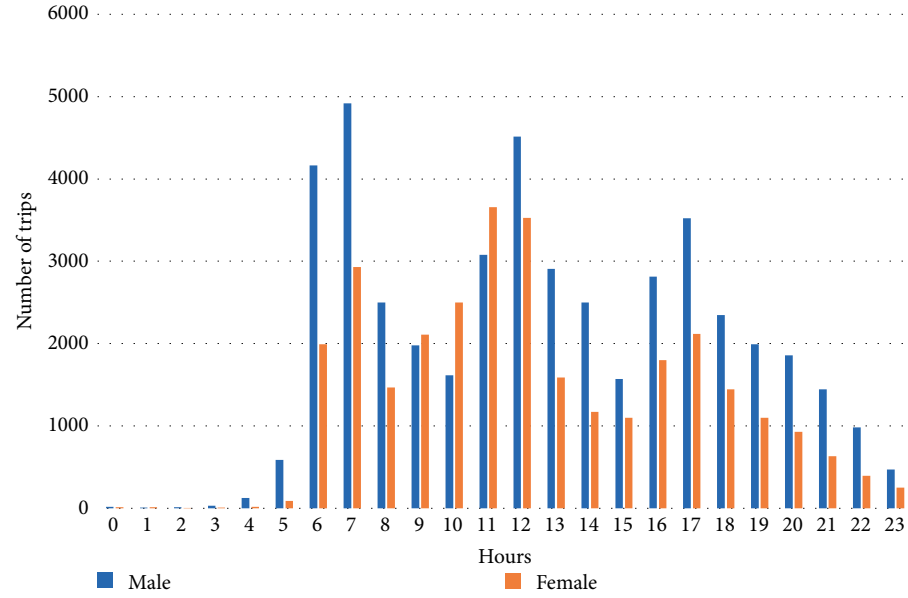

Figure 3: Temporal distribution of trips in a day by gender based on O/D survey.

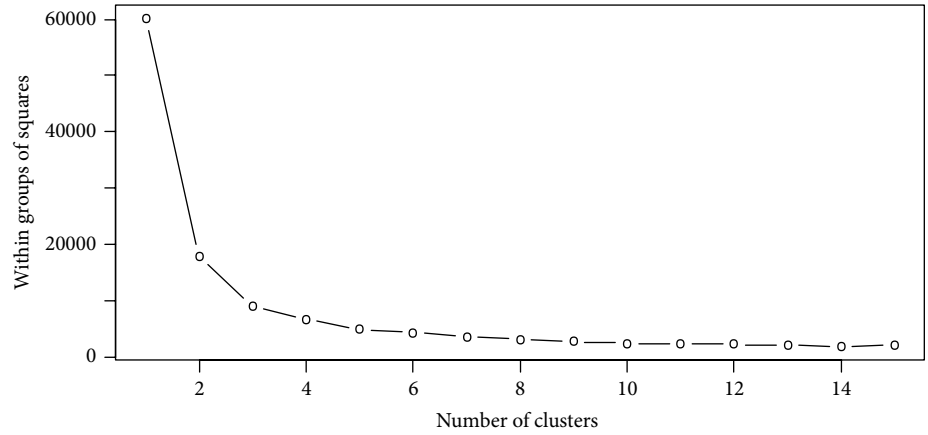

Figure 4: Scree plot obtained by K-means clustering of the data.

Table 1: Start and end time of the clusters in various methods.

Cluster no.	K-medoids		EM		K-means		Hierarchical (Ward)	
	Start	End	Start	End	Start	End	Start	End
1	2.00	6.25	2.00	5.58	2.00	6.17	2.00	6.92
2	6.27	6.75	5.67	6.55	6.25	7.25	6.58	7.42
3	6.77	7.38	6.58	7.58	7.28	8.30	6.00	7.92
4	7.42	8.33	7.60	8.83	8.33	9.50	7.50	8.92
5	8.42	10.20	8.92	10.33	9.52	10.75	8.67	10.83
6	10.25	12.75	10.42	12.75	10.83	12.75	10.00	12.75

The long travel times will dictate choosing time periods before 6.17 for morning work shifts and period of 9.50–10.75 for afternoon shifts that are shown with positive coefficient in Table 2. Also, the positive value of "travel time" coefficient shows longer travel time increases the likelihood of sooner departure time due to the importance of on-time arrival for work trips.

6. Discussion

Since the data in each cluster are different according to different clustering methods, it is not possible to use the likelihood ratio test to compare models. Thus we used the nonnested hypothesis test method, in which we examined the hypothesis

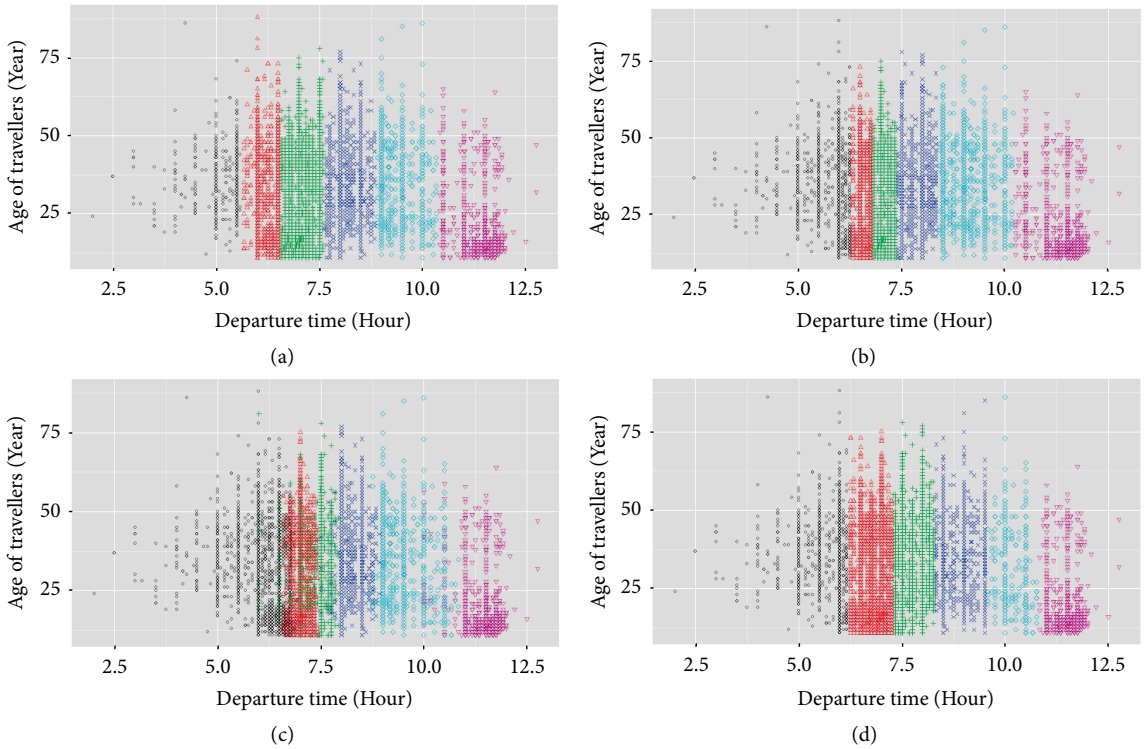

FIGURE 5: Scatterplot of Age-Departure time for various clustering methods: (a) EM. (b) K-medoids. (c) Hierarchical clustering. (d) K-means.

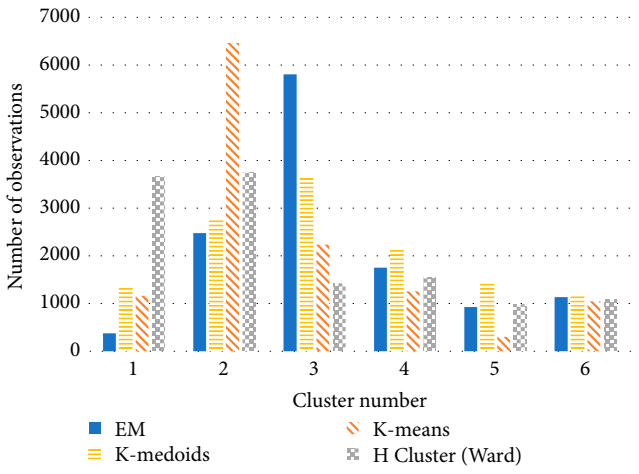

FIGURE 6: Number of each observation in various clustering methods.

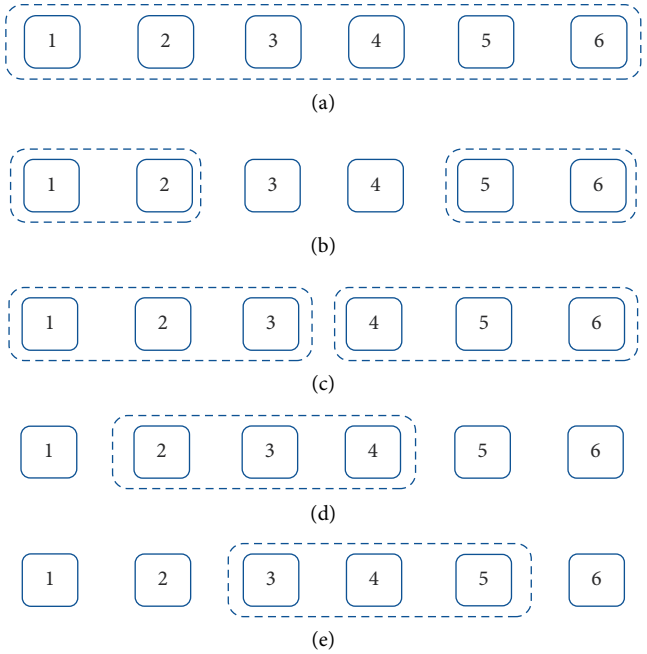

FIGURE 7: Illustration of choice sets consideration. (a) 6-membered set for all alternatives. (b) 2-membered set when traveller chooses period 1 or 6. (c) 3-membered set when traveller chooses period 2 or 5. (d) 3-membered set when traveller chooses period 3. (e). 3-membered set when traveller chooses period 4.

of the model with a smaller $\bar{\rho}^2$ value being accurate. For this purpose, the significance level is calculated via the Equation 14 [37]:

$$\text{Significance Level} = \Phi\left[-\left(-2\left(\bar{\rho}_H{}^2 - \bar{\rho}_L{}^2\right) \times LL(0) + \left(K_H - K_L\right)\right)^{1/2}\right], \quad (14)$$

where, models H and L denote the models with the higher and lower $\bar{\rho}^2$ values, respectively. Moreover, $\bar{\rho}_H{}^2$ and $\bar{\rho}_L{}^2$ are adjusted likelihood ratios for the models with higher and lower values, respectively. K_H and K_L denote the number of parameters of the H and L models, respectively. Finally, Φ stands for the normal cumulative distribution function.

Table 2: Summary of MNL modelling criteria on choice sets from various clustering methods.

Criteria	Clustering method		
	K-medoids	E.M.	K-means
Log likelihood at convergence (N = 8727)	−7700.04	−6853.41	−6166.78
Log likelihood at constant (N = 8727)	−8753.17	−8073.84	−7259.38
Log likelihood at zero (N = 8727)	−8905.19	−9171.18	−8975.74
$\bar{\rho}_0^2$	0.132	0.250	0.310
AIC	1.771	1.576	1.419
Prediction (N = 3740)	45.73%	52.34%	58.63%

Table 3: MNL model for K-means clustering method.

	Start and end time of the clusters					
	2.00–6.17	6.25–7.25	7.28–8.30	8.33–9.50	9.52–10.75	10.83–12.75
FAMILY_M					−0.15971	
					(−2.80)	
GENDER	0.37355					
	(2.90)					
JOB_G1		−1.80974		−0.53302		−0.53302
		(−10.51)		(−5.23)		(−5.23)
JOB_G2	−0.78562	−1.80426		−0.24251		
	(−5.22)	(−9.70)		(−2.10)		
JOB_G3	−0.58602	−2.89328				−0.58602
	(−4.02)	(−14.96)				(−4.02)
EDUCATED		0.28489				
		(3.89)				
CAR_OWN	−0.33386					
	(−3.61)					
SC_ MODE				−1.66956		
				(−2.26)		
PR_ MODE	−0.34566				−0.34566	
	(−3.80)				(−3.80)	
PU_ MODE			0.17971			
			(2.66)			
AC_ MODE	−0.56853					0.54963
	(−3.98)					(2.19)
AGE18P		0.35969				2.14911
		(2.12)				(8.43)
TRAVEL_T	1.55774				0.62063	
	(12.58)				(2.97)	
Constant		3.17750	0.86423	1.09339	0.07704	0.35886
		(14.04)	(5.22)	(6.20)	(0.25)	(1.60)

Table 4 illustrates the results of different hypotheses test. Following this test, the model resulted from K-means clustering is significantly better than the model produced by other methods.

As 70% of the data was used for modelling, the remaining 30% was used to validate the modelling outcomes and examine the performance of models. In other words, we used 8727 records of the data for modelling and 3740 records for validating the models. For validation of the models and determining the predictability, the observed selected choice should be compared to the alternative selected by the model, and the degree of compliance should be determined using the following measure [35]:

$$\text{Count } R^2 = \frac{\text{Number of Correct Prediction}}{n}. \quad (15)$$

We adopted two approaches to identify the model's selected alternative. The first assumed the choice with the most probability to be the chosen alternative, whereas the second determines the chosen choice by using Monte Carlo simulation

Table 4: Nonnested hypothesis test method for model comparison.

Hypothesis	Significance level	Result
$H_0 : EMmodel \succ Kmeansmodel$	$\Phi(-32.86) \cong 0$	Reject H_0
$H_0 : Kmedoidsmodel \succ Kmeansmodel$	$\Phi(-56.27) \cong 0$	Reject H_0
$H_0 : Kmedoidsmodel \succ EMmodel$	$\Phi(-45.77) \cong 0$	Reject H_0

Table 5: Different criteria for various models.

Criteria		Without model	K-medoids model	EM Model	K-means model
$\bar{\rho}_0^2$		—	0.132	0.250	0.310
AIC		—	1.771	1.576	1.419
Count R^2	Max prob.	36.3%	51.2%	57.9%	64.9%
	Monte Carlo	36.3%	45.7%	52.3%	58.6%

technique. To calculate the measure as mentioned above in the first approach, first the probability of each alternative is determined based on the models, and then the choice with maximum likelihood is assumed to be the predicted alternative and is compared to the observed choice. In the second approach, we calculated probabilities and used Monte Carlo simulation to determine the selected alternative; then we compared it to the observed choice. Monte Carlo simulation is an iterative process, and we presented the mean of various iterations in Table 5 for different models. On the other hand, if the model does not exist and we assume that the probabilities of choices in each choice set are equal, the correct prediction expected value will be 36.2% (of the 3740 data records, 700 records belong to the 2-choice set, and 3040 records belong to the 3-choice set).

7. Conclusions

The present research was carried out with the aid of discrete choice models to develop a DTC model for work trips. To this end, we employed different clustering methods for time discretization, and we compared results of MNL modelling for them. The clustering methods used in this research include the K-means, K-medoids, EM, and Hierarchical.

Although time discretization is highly significant in these models and forms the choice set, previous studies were conducted based on expert opinions on discretization, and no systematic method was proposed. The present research is an attempt to examine the performances of different clustering methods in time discretization and their effects on MNL modelling.

In MNL modelling, due to the nature of time, the choice set was designed in a way that each choice was competing with its adjoining alternatives. We compared this approach in the choice set generation to an alternative set of all available members and the higher effectiveness of the proposed method was demonstrated.

Results of different clustering methods indicated that the hierarchical method was not suitable for MNL modelling. After building the MNL models based on results of the clustering methods, some criteria were used to compare the models. Results indicated that the MNL model based on K-means results displays a better fit than other models and a higher prediction power. The prediction potential of the MNL model based on the K-means clustering method was 64.9% and 58.6% with the first and second approaches, respectively. The comparison between these values and 36.3% (the correct prediction expected value without a model) reveals the effectiveness of this model. In using the developed models, it is highly critical to consider the stability of clusters in the time horizon which depends on the durability of activity patterns and starting times of activities.

Data Availability

The departure time and socioeconomic data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there they have no conflicts of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgments

We thank M. Ahmadinejad (associate professor, Iran University of Science and Technology) for comments that greatly improved the manuscript. We would also like to show our gratitude to "anonymous" reviewers for their so-called insights. We are also immensely grateful to Francesco Viti for his comments on an earlier version of the manuscript.

Supplementary Materials

We provided supplementary tables, two of which present the coefficient of utility function in MNL modelling for K-medoids and E.M. clustering methods and third one describes the variables that used in MNL modelling. Table S1. MNL model for K-medoids clustering method. Table S2. MNL model for EM clustering method. Table S3. Description of variables that used in MNL models (*Supplementary Materials*).

References

[1] E. J. Van de Kaa, "Extended prospect theory: findings on choice behaviour from economics and the behavioural sciences and their relevance for travel behaviour," 2008.

[2] M. Fujita, S. Yamada, and S. Murakami, "Time coefficient estimation for hourly origin-destination demand from observed link flow based on semidynamic traffic assignment," *Journal of Advanced Transportation*, vol. 2017, Article ID 6495861, 14 pages, 2017.

[3] S. Afandizadeh, M. Yadak, and N. Kalantari, "Simultaneous determination of optimal toll locations and toll levels in cordon-based congestion pricing problem (case study of Mashhad city)," *International Journal of Civil Engineering*, vol. 9, no. 1, p. 33, 2011.

[4] C. R. Bhat and J. L. Steed, "A continuous-time model of departure time choice for urban shopping trips," *Transportation Research Part B: Methodological*, vol. 36, no. 3, pp. 207–224, 2002.

[5] S. Y. He, "Does flexitime affect choice of departure time for morning home-based commuting trips? Evidence from two regions in California" *Transport Policy*, vol. 25, pp. 210–221, 2013.

[6] Y. Lim and B. Heydecker, "Dynamic departure time and stochastic user equilibrium assignment," *Transportation Research Part B: Methodological*, vol. 39, no. 2, pp. 97–118, 2005.

[7] K. M. Nurul Habib, Day N. Miller, and E. J. Miller, "An investigation of commuting trip timing and mode choice in the greater toronto area: application of a joint discrete-continuous model," *Transportation Research Part A: Policy and Practice*, vol. 43, no. 7, pp. 639–653, 2009.

[8] M. Feil, M. Balmer, and K. W. Axhausen, "New approaches to generating comprehensive all-day activity-travel schedules," *Arbeitsberichte Verkehrs-und Raumplanung*, vol. 575, 2009.

[9] M. Adnan, "Linking macro-level dynamic network loading models with scheduling of individual's daily activity-travel pattern chapters," 2010.

[10] G. Cantelmo and F. Viti, "Incorporating activity duration and scheduling utility into equilibrium-based dynamic traffic assignment," *Transportation Research Part B: Methodological*, 2018.

[11] K. A. Small, "The scheduling of consumer activities: work trips," *The American Economic Review*, vol. 72, no. 3, pp. 467–479, 1982.

[12] L. Biggiero, E. Cascetta, and A. Nuzzolo, "Analysis and modelling of commuters departure time and route choices in urban networks," in *Sixth World Conference on Transportation Research*, 1992.

[13] K. A. Small, "A discrete choice model for ordered alternatives," *Econometrica: Journal of the Econometric Society*, vol. 55, no. 2, pp. 409–424, 1987.

[14] C. Hendrickson and E. Plank, "The flexibility of departure times for work trips," *Transportation Research Part A: General*, vol. 18, no. 1, pp. 25–36, 1984.

[15] A. T. H. Chin, "Influences on commuter trip departure time decisions in Singapore," *Transportation Research Part A: General*, vol. 24, no. 5, pp. 321–333, 1990.

[16] T. Sumi, Y. Matsumoto, and Y. Miyaki, "Departure time and route choice of commuters on mass transit systems," *Transportation Research Part B: Methodological*, vol. 24, no. 4, pp. 247–262, 1990.

[17] S. Afandizadeh, S. Zahabi, and N. Kalantari, "Estimating the parameters of logit model using simulated annealing algorithm: case study of mode choice modeling of Isfahan," *International Journal of Civil Engineering*, vol. 8, no. 1, pp. 68–78, 2010.

[18] F. Mannering, S.-G. Kim, W. Barfield, and L. Ng, "Statistical analysis of commuters' route, mode, and departure time flexibility," *Transportation Research Part C: Emerging Technologies*, vol. 2, no. 1, pp. 35–47, 1994.

[19] D. Ettema and H. Timmermans, "Modeling departure time choice in the context of activity scheduling behavior," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1831, no. 1, pp. 39–46, 2003.

[20] D. Ettema, G. Tamminga, H. Timmermans, and T. Arentze, "A micro-simulation model system of departure time using a perception updating model under travel time uncertainty," *Transportation Research Part A: Policy and Practice*, vol. 39, no. 4, pp. 325–344, 2005.

[21] C. R. Bhat, "Accommodating flexible substitution patterns in multi-dimensional choice modeling: formulation and application to travel mode and departure time choice," *Transportation Research Part B: Methodological*, vol. 32, no. 7, pp. 455–466, 1998.

[22] C. R. Bhat, "Analysis of travel mode and departure time choice for urban shopping trips," *Transportation Research Part B: Methodological*, vol. 32, no. 6, pp. 361–371, 1998.

[23] A. Papola, "Some developments on the cross-nested logit model," *Transportation Research Part B: Methodological*, vol. 38, no. 9, pp. 833–851, 2004.

[24] J. D. Lemp, K. M. Kockelman, and P. Damien, "The continuous cross-nested logit model: formulation and application for departure time choice," *Transportation Research Part B: Methodological*, vol. 44, no. 5, pp. 646–661, 2010.

[25] J. L. Steed and C. R. Bhat, "On modeling departure-time choice for home-based social/recreational and shopping trips," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1706, no. 1, pp. 152–159, 2000.

[26] R.-C. Jou, R. Kitamura, M.-C. Weng, and C.-C. Chen, "Dynamic commuter departure time choice under uncertainty," *Transportation Research Part A: Policy and Practice*, vol. 42, no. 5, pp. 774–783, 2008.

[27] E. Ben-Elia, M. Bierlaire, and D. A. Ettema, "Behavioral departure time choice model with latent arrival time preference and rewards for peak-hour avoidance," *European Transport Conference 2010*, United Kingdom, 2010.

[28] A. Sasic and K. N. Habib, "Modelling departure time choices by a heteroskedastic generalized logit (Het-GenL) model: an investigation on home-based commuting trips in the Greater Toronto and Hamilton Area (GTHA)," *Transportation Research Part A: Policy and Practice*, vol. 50, pp. 15–32, 2013.

[29] D. Ettema, O. Ashiru, and J. W. Polak, "Modeling timing and duration of activities and trips in response to road-pricing policies," *Transportation Research Record*, vol. 1894, no. 1, pp. 1–10, 2004.

[30] Langlois G. Goulet, H. N. Koutsopoulos, and J. Zhao, "Inferring patterns in the multi-week activity sequences of public transport users," *Transportation Research Part C: Emerging Technologies*, vol. 64, pp. 1–16, 2016.

[31] H. Chen, C. Yang, and X. Xu, "Clustering vehicle temporal and spatial travel behavior using license plate recognition data," *Journal of Advanced Transportation*, vol. 2017, pp. 1–14, 2017.

[32] Y. K. Wong and W. L. Woon, "An iterative approach to enhanced traffic signal optimization," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2885–2890, 2008.

[33] K. Train, *Discrete Choice Methods with Simulation*, Cambridge University Press, 2003.

[34] W. H. Greene, "NLOGIT version 5.0: reference guide. Econometric software Inc," Plainview, NY, 2012.

[35] J. Han, M. Kamber, and J. Pei, *Data mining: Concepts and Techniques*, Elsevier, 2011.

[36] R Core Team, "R: a language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2015, https://www.R-project.org/.

[37] M. E. B. Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory And Application To Predict Travel Demand*, MIT Press, 1985.

[38] W. H. Greene and D. A. Hensher, *Modeling Ordered Choices: A primer*, Cambridge University Press, 2010.